Automated Generation of Cognitive Ontology via Web Text-Mining

Richard Gao, Thomas Donoghue, Bradley Voytek. Department of Cognitive Science. University of California, San Diego

VOYTEKlab

UC San Diego

Summary

Goal: to map out the ontology of "cognitive processes", based on how these terms are used in publications & compare how cognitive scientists and neuroscientists conceptualize them differently.

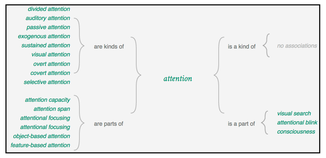
Data: PubMed abstracts and CogSci Proceedings were scraped using Python to extract term frequency and term-term co-occurrence of cognitive terms from the Cognitive Atlas (Poldrack et al, 2011).

Analysis: term frequency distribution were compared for cognitive science and neuroscience publications, and co-occurrence was submitted to hierarchical clustering to generate "empirical ontology".

Results: sensible ontologies were generated, with some (un)surprising deviations between neuroscience and cognitive science. Cognitive science tends to study more abstracts concepts and have a less diverse range of topics than neuroscience. This tool may serve as a hypothesis-generation mechanism, or semi-automated curation of cognitive ontology.

1. Motivation

Question 1: What is the ontological structure of cognitive processes?



Cognitive processes are not stand-alone operations in the mind. Can we map out how they are related to each other in an automated way?

(left) curation for "attention" in Cognitive Atlas.

Question 2: Does cognitive science inform neuroscience, and vice versa?

The history of cognitive science is rooted in computationalism, which has been an useful but perhaps outdated metaphor for the brain. When a cognitive scientist and a neuroscientist study a cognitive process, like attention, are they looking at the same thing, and in similar ways?

2. Method

Data: 800 cognitive terms from the Cognitive Atlas are used as seed terms. 4 datasets were gathered:

> CogSci: abstracts from CogSci proceedings (2008-2016) were scraped for occurrences of individual terms, and co-occurrence of terms.

> PubMed {Cogs, Neu, NeuMet}: terms were searched in conjunction with 1 of 3 sets of "base phrases", and the number of abstracts returned were recorded. e.g. attention AND memory AND (cognitive OR cognition)

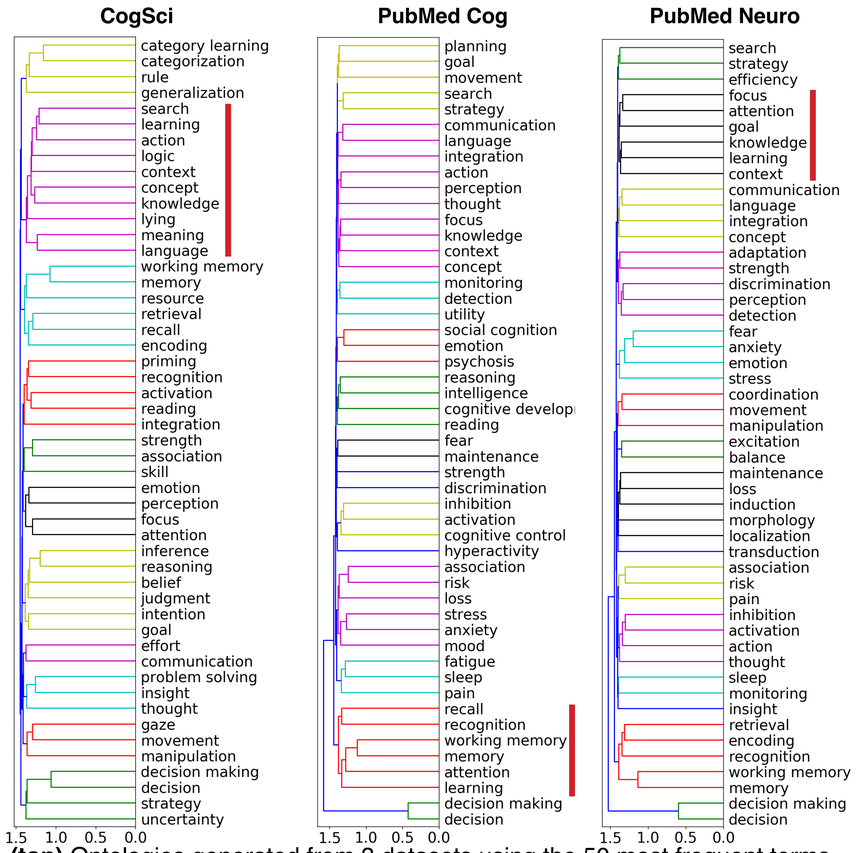
Analysis: Term frequency was normalized by the total number hits for that dataset, and co-occurrence matrix was converted to similarity via the Jaccard Index:

$$J(A,B) = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Term frequency was compared between the 4 datasets; ontology was automatically generated via hierarchical clustering.

Data and code: **github.com/voytekresearch/IdentityCrisis**

3. Cognitive Ontology



(top) Ontologies generated from 3 datasets using the 50 most frequent terms. Red bar highlights cluster containing "**learning**".

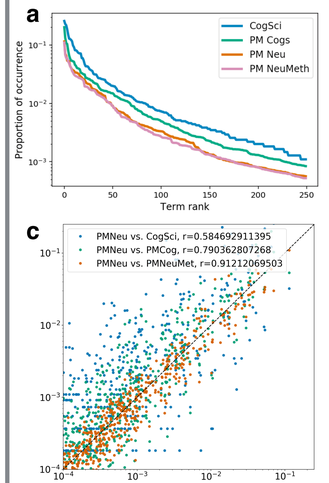
Ontology provides quick insight on relationship between processes: Hierarchical clustering of term distance matrix groups highly related terms together, forming a kind of semantic structure. Qualitatively, we find sensible clusters in each dataset, while dataset-specific differences also exist. As a whole, these results broadly survey the literature, serve as a sanity check on how neuroscientists and cognitive scientists study these processes, and reveal field-specific differences in conceptualization of these terms.

CogSci	Jaccard Index	PM CS		PM Neu	
sadness	0.167	anxiety	0.071	anxiety	0.152
happiness	0.100	extinction	0.064	extinction	0.129
shame	0.083	emotion	0.044	emotion	0.075
emotion recognition	0.063	sadness	0.038	stress	0.046
prosody	0.057	pain	0.036	consolidation	0.043
intentionality	0.053	happiness	0.031	learning	0.037
facial expression	0.053	stress	0.029	memory	0.036
stereotypes	0.048	recognition	0.029	arousal	0.032

Term similarity query: From the distance matrix, we can query for a specific term and find terms most similar to it (frequently co-occurring). This places the individual terms in the larger context of the literature.

(left) 8 closest terms to **"fear"** in 3 datasets. Note CogSci differs drastically.

4. Term Frequency Comparison



b

CogSci		PM Cog		PM Neuro		PM Neuro Meth	
learning	0.257	memory	0.201	activation	0.114	activation	0.11
search	0.228	risk	0.131	learning	0.071	detection	0.068
action	0.218	attention	0.108	memory	0.065	memory	0.06
language	0.185	learning	0.104	loss	0.054	risk	0.059
logic	0.151	association	0.081	action	0.053	attention	0.052
knowledge	0.128	anxiety	0.066	inhibition	0.052	monitoring	0.048
concept	0.126	stress	0.059	attention	0.044	sleep	0.043
memory	0.122	loss	0.054	risk	0.041	association	0.041
context	0.116	activation	0.053	perception	0.038	movement	0.036
decision	0.098	knowledge	0.052	detection	0.037	loss	0.034
attention	0.091	language	0.051	knowledge	0.036	action	0.031
reasoning	0.081	context	0.046	association	0.034	learning	0.03
judgment	0.081	working memory	0.044	context	0.034	perception	0.029
focus	0.079	focus	0.04	stress	0.033	inhibition	0.029
lying	0.074	perception	0.04	movement	0.032	knowledge	0.028
inference	0.072	recognition	0.039	recognition	0.032	focus	0.027
perception	0.072	mood	0.038	induction	0.03	localization	0.027
meaning	0.071	action	0.034	integration	0.029	context	0.027
movement	0.06	pain	0.033	encoding	0.028	recognition	0.026
goal	0.059	sleep	0.032	thought	0.026	language	0.025

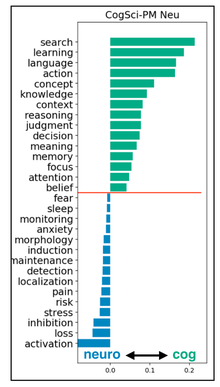
a. Proportion of occurrence of top 250 terms in each dataset.
b. Top 20 terms in each dataset and their proportion of occurrence. Green: unique terms in the top-20 of each dataset; red: neuro only terms.
c. Pair-wise similarity between PubMed Neuro dataset and other 3 datasets.

Cognitive Science uses broader terms, and more frequently: Frequently used terms in Cognitive Science datasets are used much more commonly (e.g. learning, memory), and they are often high-level concepts for cognitive processes. Specific term usage in cognitive science deviates greatly from in neuroscience. This suggests that cognitive scientists and neuroscientists have different focus.

5. Conclusion

Cognitive science informs neuroscience? If cognitive theories are to make a broad impact on our investigation of the neural bases of cognitive processes (Frank & Badre, 2015), a convergence of ontology in these literatures should follow.

(right) Highest term usage difference for 2 datasets.



Tool for automating meta-analysis: Cognitive processes are often related to one another (Poldrack & Yarkoni, 2016), and we define them in relation to each other in the literature. This tool may serve as an overview or semi-automated curation.

(left) curation for "**addiction**" in the Cognitive Atlas, compared to the generated ontology.

poster reprint: **rigao@ucsd.edu**

@_rdgao

NSERC CRSNG

Alfred P. Sloan Foundation